Towards a systematic classification of protein folds

Lindgård, Per-Anker; Bohr, Henrik

Published in:
Physical Review E. Statistical, Nonlinear, and Soft Matter Physics

Link to article, DOI:
10.1103/PhysRevE.56.4497

Publication date:
1997

Document Version
Publisher's PDF, also known as Version of record

Towards a systematic classification of protein folds

Per-Anker Lindgård
Department of Condensed Matter Physics, Risø National Laboratory, DK-4000 Roskilde, Denmark

Henrik Bohr
Center for Biological Sequence Analysis, Department of Physical Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark

(Received 26 November 1996; revised manuscript received 28 May 1997)

A lattice model Hamiltonian is suggested for protein structures that can explain the division into structural fold classes during the folding process. Proteins are described by chains of secondary structure elements, with the hinges in between being the important degrees of freedom. The protein structures are given a unique name, which simultaneously represent a linear string of physical coupling constants describing hinge spin interactions. We have defined a metric and a precise distance measure between the fold classes. An automated procedure is constructed in which any protein structure in the usual protein data base coordinate format can be transformed into the proposed chain representation. Taking into account hydrophobic forces we have found a mechanism for the formation of domains with a unique fold containing predicted magic numbers {4,6,9,12,16,18,...} of secondary structures and multiples of these domains. It is shown that the same magic numbers are robust and occur as well for packing on other nonclosed packed lattices. We have performed a statistical analysis of available protein structures and found agreement with the predicted preferred abundances of proteins with a predicted magic number of secondary structures. Thermodynamic arguments for the increased abundance and a phase diagram for the folding scenario are given. This includes an intermediate high symmetry phase, the *parent* structures, between the *molten globule* and the *native* states. We have made an exhaustive enumeration of dense lattice animals on a cubic lattice for acceptance number $Z=4$ and $Z=5$ up to 36 vertices. [S1063-651X(97)04909-X]

PACS number(s): 87.10.+e, 05.50.+q, 05.70.Ln

I. INTRODUCTION

In the past 50 years large databases of protein sequences and protein structures have been building up at an exponential rate [1]. And, as in the case of, for example, atomic elements or isotope tables, it is natural to ask for some classification that can group the proteins into related families other than those that arise from homology analysis of the sequence of amino acids in the polypeptide chain. What we have in mind here is a kind of atomistic taxonomy, where the proteins are grouped according to the number of typical elements.

In the case of the nuclear isotopes the grouping in particularly stable, closed shells of nucleons came rather late historically, since it was not obvious that an independent-particle description would make sense in the nuclear interaction picture, and yet magic numbers came out of a fairly simple single-particle force potential. This led to a predicted predominance of abundance of nuclei at magic numbers of nucleons, in agreement with empirical data. Likewise for our microbiology case we shall show that magic numbers for the stability in the packing of protein structure elements are revealed in a calculation based on a simple hydrophobic force field model. Proteins appear to be packed like closed ''shells'' of all connected secondary structure elements. The purpose of this paper is to provide a paradigm that allows classification of the proteins in structurally defined families.

Let us briefly list some pertinent features of protein structures and the folding process. Excellent reviews can be found in [2] and more details about the experimental facts and unsolved questions are given by, e.g., Jaenicke [3,4] and from a theoretical point of view by Finkelstein and Ptitsyn [5,6] and Wolynes [7].

Proteins are found to be highly hierarchically structured. Pauling and co-workers [8] were the first to emphasize that the final, so-called *native* structure of proteins consists of two dominant kinds of building blocks, the α helices and the β sheets. These are called secondary structures. Later additional, somewhat less characteristic structural elements were proposed (i.e., inverse turns and Ω loops, etc. [9]). A characteristic feature of proteins is that their observed structures are densely folded in a complex manner of secondary structures and intervening irregular loops [9]. These further form tertiary structures, which are composed of characteristic domains with a special fold, which are made up of typically tens of secondary structures. The domains further self-organize into quartinary structures consisting of several domains. Dense folding intermediates are observed before reaching the unique closed packed state [10].

In aqueous solutions most proteins fold after various intermediate stages [3,4] into closely packed globules, which neither dissolve nor phase separate, as most polymers would do. Dill [11] derived a thermodynamic theory for these and showed they should have a tendency to fold into lumps of specific size. A main reason for this is the action of the hydrophobic and hydrophilic forces, which are unspecific interface-tension-like forces [12,13]. Yet, a protein with a specific amino acid chain folds, paradoxically [14] in a matter of seconds, to a particular *fold*, according to information that must be provided via the underlying linear information

represented by the specific sequence of amino acids. Furthermore proteins seem to have predominant lengths of the chains. Berman *et al.* [15] have made a statistical study of known proteins and have found that the distribution has characteristic peaks near multiples of chain lengths of 125 amino acids. The total length may go up to a few thousand. About 400 distinct structures are known [1] from x-ray crystallography for such domains, but only for proteins that form crystalline structures, i.e., not in the more relevant environment, the natural solution with salty water. These are grouped into a few hundred recognized fold classes. Less detailed structural information in the solutions or *in vivo* are available from NMR and circular dichroism studies. In total ~ 4000 structures have so far been determined [16], however, several appear to be closely related. On the other hand, well over hundreds of thousands of proteins have had their sequence determined [17]. It is of great interest (1) to be able to predict a structure from the sequence, (2) to be able to classify the possible structures that can exist, and (3) to understand why certain structures seem to be particularly abundant. The aim of this work is to propose a schematic framework for the description of the folding of secondary structures into domains of proteins and discuss their abundance.

First, consider the simpler crystalline classes of structures. Group theory tells us that there are only 230 different classes in three dimensions. Many materials assume before they melt, in spite of the possible diversity, a single open structure, the body centered cubic structure bcc, which is stabilized by entropy; see, e.g., [18]. This is called the *parent phase*. At lower temperature the structure transforms by a so-called Martensitic transformation to more closed packed structures with generally ''triangular'' coordination between the constituents. There can be several such possibilities, hcp, fcc, dhcp 9-R, 18-R, . . . , however, all are resulting from the single ''parent'' bcc phase [19]. The observed, irregular protein structures may correspond to such complicated ground state configurations, which are the result of the competition between all relevant forces. It is too complicated to make a classification for these. However, we demonstrate that it is possible that the protein also first forms a high symmetry, dense *parent phase* from which the actually observed, still more closely packed structures are obtained by ''twisting.'' This is in order to satisfy the short ranged forces between the secondary elements. We shall postulate that a parent phase is an important intermediate phase in the folding process. By this and by considering a rather general three-dimensional (3D) structural model, our approach differs from the previously forwarded ideas to simplify the description of protein into ''folding patterns'' or ''crude structures;'' see, e.g., Finkelstein and Ptitsyn [5]. Unfortunately the experimental structural information, at present, is rather scarce on the intermediate phase [20,21]. However, the presence of intermediate phases and folding steps is a generic feature of the folding process [3,4] and some steps are described as rate limiting.

In the course of this work we numerically evaluate and exhaustively count graphs on a simple cubic lattice. This is of general applicability in a class of statistical problems. Our counts are extended to larger lattice animals than hitherto considered. Our results agree exactly with those of Chan and Dill [22,23], where overlap exists. Chan and Dill further did a graph theoretical analysis, which is of relevance for the present case as well.

The structure of this paper is as follows. First, we present the motivation and prerequisites for setting up a simplified model, still containing the pertinent physics and symmetry. Then we formulate a homology measure, which allows a systematic naming of structures and a distance measure. Using the model we find numerical evidence for magic numbers. We perform a statistics of the abundance of secondary structures and of proteins with a certain number of secondary structures. We motivate the magic numbers geometrically. Finally in the last section we make a thermodynamic theory for our model that formalizes the discussed folding scenarios and gives a thermodynamic motivation for higher abundance at the magic numbers.

II. CLASSIFICATION OF PROTEINS INTO FOLD CLASSES

It is important to understand how the proteins can find their fold without trying all the statistically possible options. It is generally assumed that the information is coded linearly in terms of the amino acid sequence, giving rise to a natural tendency for the backbone to fold correctly and fast. An unsolved problem is to demonstrate how the sequence information (which determines foremost the short range forces along the backbone and only more indirectly the interactions between distant parts of the chain) is sufficient to do this. It is our thesis that the nonlocal forces between distant sections of the proteins come in at a late stage, only providing the final optimization, and the observed complex irregular and twisted patterns. The hydrophilic and hydrophobic forces against the aqueous solution are supposed to be the main driving forces in condensing the proteins from the extended state. The protein chain has about 50% hydrophobic and hydrophilic residues distributed seemingly at random along the chain. An extended chain is, therefore, clearly unfavorable. The optimum is a condensed phase with a minimal surface, which allows most of the hydrophilic residues to be buried. However, it is not possible for the unspecific hydrophobic forces to define a *specific fold* when the system is in an unfolded state. A fold means [24–30] a particular structural topology that a protein domain can assume in its native state.

Proteins appear to belong to families, like plants, with specific characteristics. The families contain many variants. von Linné [31] in the 18th century succeeded in the field of botany to identify the important classification parameters. He solved the difficult *homology* problem defining when plants are *the same* without being *identical*, and when they belong to the same class or not. It gives a systematic, although not ''natural'' classification from a functional point of view. Here we suggest that the dense fold patterns for proteins may form the basis for a classification, and we shall identify a class of similar folds with a family, as did Chothia [24] (and with the qualifications mentioned that the fold classes need not be the natural families). By devising a local projection scheme for systematizing the protein fold on a lattice we propose an effective cut through the homology problem. The results were briefly discussed previously [32]. Such a schematic structure is a kind of symmetry indicator [33], which is useful in statistical analysis of the fold problem. It is well

known that a global measure for ''similar'' folds using the root-mean-square measure (rms) for the coordinates of the backbones is too strict, and indeed vastly misleading; see, e.g., [34]. If just one secondary element is slightly rotated, the rms can become very large; this is not expedient. Other measures, for example, local distance measures, have also been proposed and used [1]. In traditional classification in physics, as in the periodic table or in the crystal groups, a certain capaciousness in the homology concept is neither needed nor warranted. In the protein folding case, as in botany, it is. Yet the final classification criteria must be unique.

Similar simplifications with idealized elements have previously been proposed by Murzin and Finkelstein [35] for describing the domains of α helices. They considered the α helices as cylinders and considered a close packing of these on edges of polyhedra with triangular faces. They demonstrated a high degree of coordination of the possible and the observed structures, except for bundles of larger numbers of long helices, which seem to align more in parallel. It is interesting to note that their structures in all cases can be regarded as twisted structures of a simple parallel bundle. Their work describes the number of distinct twists. In the above crystal analogy, they classify some of the possible closed packed structures belonging to a single ''cubic'' parent phase. The polyhedron method has the drawback that it does not work for β sheets. However, our cubic representation describes equally well the β sheets and the β sandwiches, which are schematized in a different representation by Finkelstein and Reva [36].

Recently, even more schematized compact lattice models for late stages of protein folding in terms of a chain of interacting beads (monomers) have been intensively studied [37–44]. Secondary structures are very schematically modeled as sequences of monomers with a persistence length of two or more beads, usually on a $3\times3\times3$-bead cube. The model proteins are supposed to be refolding and forming the secondary structures at the compact folding stage in a search for the minimum of strong interchain interactions (represented by two or more attractive or repulsive beads, randomly distributed), or for a state of ''minimum frustration'' as discussed by Wolynes [7]. This approach is very different from the present case. It is an interesting and useful model in its own right in particular for heteropolymers. It is focused on the difficult problem of describing a frustrated search for the optimum in a rugged energy landscape. That is undoubtedly very relevant for proteins too, however, in our model we take almost the opposite view and take maximum advantage of proteins' proven ability to form secondary structures at an early stage.

III. A MODEL HAMILTONIAN FOR PROTEIN FOLDING

In the following we shall construct a minimal model for protein folding in order to establish a vocabulary and a language in which the structures can be described and subsequently classified. Summarizing the review of 20 years of protein folding research Jaenicke [4] concludes that the process can be described as a multiple pathway of sequential folding with roughly three steps: (1) very fast early events, (2) middle events with local shuffling into tertiary structure, and finally (3) the late events forming the chemical bonds (disulfide bridges, etc.). All three steps can be assisted by other proteins (so-called chaperones) [4]. We shall model these observations with special emphasis on the second stage.

Experimentally the helix structures are usually seen to form in the very early stages of the folding process [4], although not without exceptions, which indicates that in some cases the final form of the secondary structures is obtained in concert with the overall folding [20,21]. The helices are typically between 4 and 12 amino acids long (see Fig. 6), which in fact can be understood on the basis of a simple random copolymer model [5]. At relatively high temperatures, i.e., above or at the *molten globule state* (which is an operational term for a rather dense state with pronounced secondary structure; see, e.g., [2], p. 265) we assume, in agreement with Jaenicke's conclusion, that the protein is substructured according to the underlying amino-acid letter code, into two groups of secondary structures, as can be seen in the well-known ribbon representations of proteins [45]. One set, which we denote by capital letters $A,B,C,\ldots,$ represents the described helices [46] and also potential strands for the formation of β sheets. Strictly speaking, the latter cannot be well described at this temperature since their stabilization probably requires also the forces between different parts of the protein and not just forces along the backbone. Yet, the β strands need to be folded into the correct relative position in space. These elements are assumed to be approximately linear with a well defined start and end point (amino acid). The secondary elements can with quite high confidence be predicted from the linear sequence information based on the DSSP (definition of secondary structures of proteins) algorithm [47]. The second group consists of the remaining connecting pieces of the protein, the irregular loops (which have an average length of 4 residues, Fig. 6). These can be replaced by the straight connection line, $a,b,c,\ldots$ between two consecutive secondary elements of the first group. Then all elements can be considered straight. Two elements are connected by a ''hinge,'' which is characterized by a direction in space, perpendicular to the plane in which the two joining elements can rotate. The position and action of the hinge are in principle determined by the underlying amino acid sequence; however, the code is yet to be found by statistical analysis. Using a spin $\mathcal{S}_i$ for this description we can define both the direction and the sense of the bend between the two elements. We then make the crucial, simplifying assumption that each element is sufficiently rigid to define the relative optimum direction of the spins attached to the ends of the element.

Thus the protein is schematized as the sequence of secondary structures and connections with preferential bending forces acting between them

$$a\mathcal{S}_1A\mathcal{S}_2b\mathcal{S}_3B\mathcal{S}_4c\mathcal{S}_5C\mathcal{S}_6d,\ldots\ . \tag{1}$$

It is at this level that we shall attempt to classify the various protein foldings. We are now ready to formalize the model in order to be able to make computer simulations and predictions of fold classes. This scheme is not simply a lattice model, and in principle it can be made general with arbitrary angles and lengths. At a later stage we shall include interac-

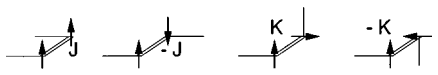

FIG. 1. Definition of the hinge spins and the hinge force parameters for secondary elements, double line. The definition is analogous for the intermediate elements, single lines. The drawing is in perspective, all angles represent 90°.

tions between the elements of the first group $A,B,C,\ldots$, in particular between the potential β-sheet elements.

A. A fold Hamiltonian with the pertinent symmetry

We need further simplification to get a practical model for fold structure formation. For a statistical description it is probably not important to allow continuous variations in the possible angles, so we assume only one allowed angle, and the value of the angle is not essential for the argument in the first stage. For ease of representation we choose this to be 90° (later also including the value 0°). Let us traverse the protein represented by Eq. (1) from left to right. Each element $P=A,B,C,\ldots$ has a direction unit vector $\hat{\mathbf{e}}^P$ along one of the axes in a Cartesian coordinate system. We remark that using simply the direction vectors makes the description independent of the lengths of the elements. This is a simplification based on the fact that the actual elements have lengths of the same order of magnitude; see Fig. 6. It is also independent of the position in space and of interactions between the elements apart from direct overlap. Similarly each element $p=a,b,c,\ldots$ is characterized by $\hat{\mathbf{e}}^p$. The structure is given by the sequence of spin vectors $\mathcal{S}_1,\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4,\ldots$. The spins have unit lengths and may each point in either of the six directions $\pm x,\pm y,\pm z$. If we consider only the 90° (and 0°) turns, a unique description for the orientation between two elements a and A with a hinge spin $\mathcal{S}_1$ (and further b joined by the hinge spin $\mathcal{S}_2$) is given by

$$\hat{\mathbf{e}}^A=\hat{\mathbf{e}}^a\times\mathcal{S}_1+(\hat{\mathbf{e}}^a\cdot\mathcal{S}_1)\hat{\mathbf{e}}^a,$$
$$\hat{\mathbf{e}}^b=\hat{\mathbf{e}}^A\times\mathcal{S}_2+(\hat{\mathbf{e}}^A\cdot\mathcal{S}_2)\hat{\mathbf{e}}^A,\ \text{etc.} \tag{2}$$

It is clear that the fold is uniquely described by the sequence and state of the ''hinge'' variables, the spins $\mathcal{S}_i$. A given sequence of spins $\mathcal{S}_i$ and the start direction $\hat{\mathbf{e}}^a$ is a rigid building prescription by which any later element direction $\hat{\mathbf{e}}^i$ is exactly determined.

However, this is too strict, and we want just to give building guidelines. For an element of group one, which may be optimally surrounded by parallel spins ($\uparrow A\uparrow$), let us say it gains an energy J if the spins are parallel, gains nothing if they are perpendicular ($\uparrow A\rightarrow$), and pays an energy $-J$ if the spins are antiparallel ($\uparrow A\downarrow$). If the spins should have a right turn we would give an energy gain K for the right turn, 0 for parallel or antiparallel, and $-K$ for the wrong, left twist. The possibilities are shown in Fig. 1. We can define similar energy conditions for elements of group two, with possibly different, and lower energy values j,k. We then form a linear chain of these energy variables, describing the preferred state of its surrounding spins, e.g.,

$$\Updownarrow j\Updownarrow K\Updownarrow k\Updownarrow(-K)\Updownarrow j\Updownarrow J\Updownarrow(-j)\Updownarrow\cdots, \tag{3}$$

where $\Updownarrow$ represents any of the possible six spin directions for the hinge spins. We notice this is a more flexible description than Eq. (1). The structure is now determined by the interaction constants sequence given in Eq. (3), as an example, as $j,K,k,-K,j,J,-j,\ldots$. This gives a unique best set of the spin variables $\mathcal{S}_1,\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4,\ldots$. From those the ground state can be constructed from Eq. (2). If that is all we want, we could just as well take all constants equal in magnitude, say equal to one, leaving just the signs. This would be a kind of interaction ''spin'' variables. However, we could also consider ''wrong'' folds and then it would be nice to have different energy parameters to give us the energy cost for that. A change in a spin ($\mathcal{S}_i$) direction at a junction i has the dramatic consequence of rotating the entire remaining pieces of the protein around this junction. We shall assume that there is no inertia and no stearic hindrance in doing so (this could in fact also be introduced in the model). Expressed in another way, we do not care how the system has arrived at any state for which we can measure the energy. This is reasonable when discussing the ground state. In order to be able to describe the energy cost for violating the optimum fold we write the argument as a Hamiltonian:

$$\mathcal{H}_{\text{hinge}}=-\sum_P\ (J_P\mathcal{S}_P\cdot\mathcal{S}_{P+1}+K_P\mathcal{S}_P\times\mathcal{S}_{P+1}\cdot\hat{\mathbf{e}}^P)$$
$$-\sum_p\ (j_p\mathcal{S}_p\cdot\mathcal{S}_{p+1}+k_p\mathcal{S}_p\times\mathcal{S}_{p+1}\cdot\hat{\mathbf{e}}^p). \tag{4}$$

We neglect the orientation of the beginning and end loops. In Eq. (4) $P=2n+1$ and $p=2n$, where the index n is running from $n=0$ to $\frac{1}{2}(N-1)$, where N is the number of elements. The constant $J_P=+J$ or $-J$ determines the energy for having the spins at the ends of a group one element P as parallel or antiparallel spins in the x, y, or z direction. The constant $K_P=+K$ or $-K$ determines the energy for having the spins perpendicular or antiperpendicular to each other (right and left thumb rule), and similar for j_P and k_P. To simplify the notation we shall sometimes write $-X=X$. We have here disregarded the cases with angle 0°, and cases with the spins along the element direction. The choice number is therefore by construction $Z=4$, which is the lattice coordination number minus two. One may start by fixing, e.g., $\mathcal{S}_1=\hat{z}$ and $\mathbf{e}^a=\hat{x}$; the rest then follows from Eq. (2). For the α helix it is rather clear that the interaction between the spins will be simply related to the number of amino acids that form the helix. For a random sequence of the interaction constants the model exhibits known folds among a wealth of other structures such as noncompact, loosely packed structures and structures that are too densely entangled in one another. One can characterize a given fold configuration uniquely by a linear string of coupling constants, which in fact is our systematic name or name for the fold class. As an example, the string $jKjKj$ is our systematic name for a so-called four-helix bundle protein: $1hmq$. It is important to note that there is rotational invariance of the constants $J,K,\ldots$, and therefore of the representation of the proteins by such constants, contrary to a vector representation. As a simple example we have shown in Fig. 1 of Ref. [32] the projection of the 4-α-helix bundle, which is denoted as $jKjKj$. The name depends

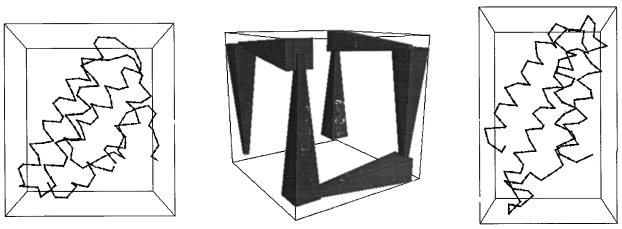

FIG. 2. Two different proteins, haemarythrin and cytuchrome b_{562} belonging to the 4α helix fold class represented as the same configuration in the computerized chain link-arrow scheme.

on the direction in which the chain is traversed, but it is invariant under rotation and translation. Some proteins are ''embellished'' by the addition of several amino acids. This focuses on the question of a definition of families and of a metric. We suggest as a measure for closeness between classes that two proteins, not necessarily of the same length, have the largest similarity if the overlap in their names is maximal (see Fig. 2).

The reduced information in the name giving the spin directions can be furnished by many amino acid sequences. This provides in fact the basis for the classification, i.e., many sequences may have the same fold.

We must also judge energy differences between good and bad folds for the same sequence. We need a simple compactness measure. On a simple cubic lattice a dense packing of a chain can be defined as one in which all vertices have the maximum number of nearest neighbors. This measure has been used earlier by Chan and Dill [22] and Camacho and Thirumalai [39] (for the bead model). Another measure of optimal packing with respect to the hydrophobic forces acting on the secondary structures, which usually have a predominant hydrophobic side, takes into account that these are to be packed as closely as possible [5]. Then we need a subset of the above classes (characterized by the interaction constants) in which the secondary elements in addition have as many parallel neighboring elements as possible.

In order to make the classification applicable to actual proteins it is important to have a unique and easy identification of the class to which any given protein belongs. Since the observed low-temperature structures are usually strongly twisted a global projection on a cubic lattice is not meaningful. We wish to devise a *local* identification method as follows. Find the unit vectors along the elements. For the loops this represents the interaction line between two connected type 1 elements. For any three consecutive unit vectors $\hat{\mathbf{e}}_0$, $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$ the condition $\hat{\mathbf{e}}_0 \cdot \hat{\mathbf{e}}_2 > 1/\sqrt{2}$ defines the interaction constant for element 1. The sequence information for this element is then reduced to one letter.

B. Distance measures between fold classes

We can define a metric on the space of folds. Firstly, two folds belong to the same class if their projected paths traced out by their backbone are identical on the 3D lattice; this is uniquely described by the string of coupling constants (e.g., $jJkK$, etc.), thus providing the name for the fold class. One of the key points in this paper is that the rather loose notion of fold classes can now be rigorously characterized by the names defined here. To see that such a characterization makes sense we can again take the example of the 4α-helix bundle fold class and inspect that the different protein domain members have roughly the same name. For example, haemarythin, tulysozyme, and cytochrome b_{256} are all given by the same name $(jKjKj)$ whereas rice cytochrome c (1CCR) from another fold class also has a different name $(jKkKj)$.

An oversimplification of the presented formalism is that α-helix and β-strand elements are not distinguished. This means that at present the helices in the helix bundle class can be replaced by β strands. This can be easily remedied by introducing special interaction parameters for the β strands. Thus for those we assign instead of $\pm J, \pm K$ the different constants, e.g., $\pm I, \pm H$, which gives four more letters in the alphabet, 12 in all, by which to write the name. The structural restrictions that β strands are close in space means that the new letters will not be randomly placed, but more likely be in close groups. A further generalization is to consider the mentioned direct move of two adjacent elements. Since this can happen for both types, we need to introduce both $\pm L$ and $\pm \not{}$, thus again adding four extra letters to the fold code, 16 in all.

The most systematic way to define a distance between fold classes is to use the difference in the names of the classes. For two names with N_1 and N_2 letters the distance $D^{\max}_{\text{sequence}}$ can be defined as

$$D^{\max}_{\text{sequence}} = N - N^{\max}_{\text{is}}, \tag{5}$$

where $N^{\max}_{\text{is}}$ is the number of letters in the maximal identical sequence (is), and $N = \max\{N_1, N_2\}$. We can also define a more average distance measure in terms of the sum of the number of matching sequences:

$$D^{\text{sum}}_{\text{sequence}} = N - \sum N_{\text{is}}. \tag{6}$$

In the 8-letter code the name of the fold classes for the 4α-helix bundle and the β-sandwich plastocyanine will have a certain overlap (due to the fact that helices and strands are counted the same) and therefore a small distance between them, while the 4-helix bundle and the TIM barrel will have a large distance between them (consistent with their great differences in size and geometry). This rough classification is useful if we are mostly interested in quantifying geometrical and topological (or morphogenetical) aspects of the structures of proteins more than their content. To include the aspect of content we must just use the above defined 12-letter code, which clearly ensures that, with the same measures, now the 4α-helix bundle and the 4β-sandwich belong to very different classes.

We would like to mention that it is of course possible to translate the interaction constant names into more phonetic and pronounceable names. One assignment with obvious mnemonic value is the replacement of J,J,K,K by b_{ack}, f_{orward}, r_{ight}, l_{eft} and $j, \overline{j}, k, \overline{k}$ by i_{nvert}, a_{dvance}, o_{ver}, u_{nder}. It shows that our interaction constants simply are ''road instructions'' for navigation in 3D space [48]. This analogy indicates in fact that the choice of the exact cubic lattice with exactly 90° turns is probably not too restrictive. With this replacement, for example, the 4α-helix bundle be-

comes instead of $jKj\overline{Kj}$ simply *irili*. There are in all six different variants with 4α-helices, namely, $jKjKj$, $kJkJk$, $jKkKj$ and those with the signs changed on the K and k parameters corresponding to a reverse fold, where the N and C termini have been interchanged. With the phonetic names these are *irili*, *obubo*, *iruri* and *iliri*, *ubobu*, *iloli*. Besides being mnemonic, they are clearly much easier to comprehend than the interaction constants, although of the same information value. Similar, highly pronouncable and structured names are found for the larger densely packed folds [49], which are far from just a random selection of the eight letters. Further it is found that the names for very long proteins (with, e.g., 35 elements and thus a 33-letter name) tend to decompose into a compound of two or more names for smaller ones (much as long words in actual languages are compounded). This is a sign of the fact that the 3D dense packing tends to favor the formation of subdomains or fold motifs.

IV. NUMERICAL CALCULATION OF THE FOLD CLASSES

The purpose of the numerical calculation is to find precisely how many densely packed configurations of a given chain can exist on the 3D regular lattice. From this number we estimate (1) the number of specific folds and (2) the total number of possible fold classes, (3) besides gaining statistical knowledge of configurations for (4) a particular number of elements and lattice sizes. The latter turns out not to be crucial since the statistics of the dense configurations converges to the correct value for larger lattices.

Using the fact that the hydrophobic forces condense the proteins and make them contain as little as 3% water [11] in the native state, we want to find all folds that are self-avoiding and densely packed. The dense packing criteria we have used is a simple count of the neighbors of end points of the elements (vertices). This does in fact represent the hydrophobic force faithfully. Firstly, it is unspecific, i.e., independent of which elements are close to each other. Secondly, it depends on the ''curvature'' of the confinement approximately as a surface tension force, i.e., the different sites are rated 3, 4, 5, and 6 for a corner, edge, face and a buried site, respectively. Only the sum counts, in agreement with the nature of the hydrophobic force. One could, in order to introduce a temperature in the problem, assign energy values for the mentioned sites. This need not be a linear weighting. If the weighting is far from linear one can form other families of proteins. For example, such that are dissolved in cell membranes. Clearly, for those the hydrophobic and hydrophillic forces act differently. Families could be imagined with higher choice number Z or other projected lattices as discussed in the next section. We have investigated the closed packed folds for the simple cubic lattice case with $Z=5$. The fact that there may be a range of different families does not invalidate our theory for the classification of globular proteins.

Let us now describe how one can calculate numerically all chain configurations in a given regular lattice setup (with a given lattice size) and for a chain with a given number of elements. For mapping out the ground state, it is most straightforward to operate directly on the element direction

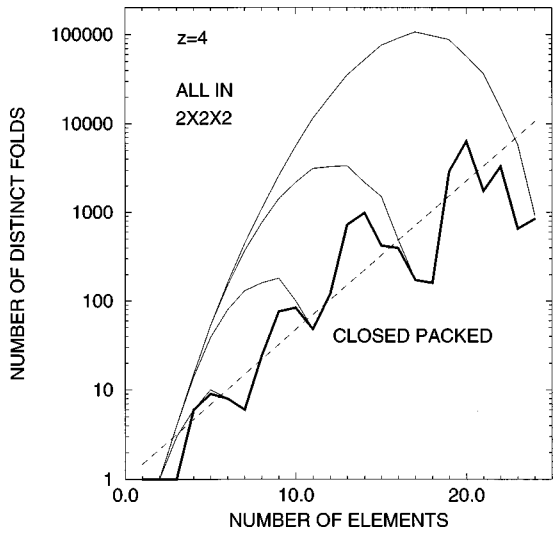

FIG. 3. Full thin line, number of distinct folds for coordination number $Z=4$, on a cubic lattice in a $2\times2\times2$ box as a function of number of elements N; the number in the smaller enclosed boxes $1\times1\times1$, $2\times1\times1$, and $2\times2\times1$ are also shown. The thin dashed line is the mean field estimate $(Z/e)^N$. The thick line shows the number of dense folds.

vectors $\hat{e}_P$ and $\hat{e}_p$. We start by placing two perpendicular elements (a,A) and their spin: $\hat{e}_1^a\mathcal{S}_1\hat{e}_2^A$. The next element direction vector $\hat{e}_3^b$ is then placed in any of the four possible directions according to the values of the interaction constant for element A: $\pm J$, corresponding to an element parallel and antiparallel to $\hat{\mathbf{e}}_1^a$, and $\pm K$ corresponding to an element perpendicular or antiperpendicular to $\hat{\mathbf{e}}_1^a$. This determines the direction of $\mathcal{S}_2$, which is not essential for the ground state calculation, since all spins follow the direction dictated by the interaction constants. However, the spins are important for the excited states since the spin flips describe the excursions from the optimum folds. All the possible positions of the third element are included as long as the element stays within the lattice box and does not collide with the previous elements. Next, we determine the allowed positions of the fourth element on the chain. We try again all the possible directions of the fourth element, use those that avoid colliding with other elements, and stay within the lattice and discard the others. This procedure is repeated until all the elements of the chain are positioned and hence we obtain a whole set of chain configurations each described by a set of coupling constants (J,K, etc.). Since distinct configurations are described by different sets of coupling constants, it is possible to sort the number of chain configurations uniquely. We vary the initial conditions so that the first two elements will be positioned over the entire lattice. The sorting by names ensures that only ''irreducible'' configurations, which cannot be brought into another by a simple symmetry operation, are counted. (Because we distinguish between the direction of traversing the chain we distinguish between reflected configurations and obtain consequently counts that are a factor of two larger than the ''bare'' dense counts.)

The process is then continued, under the constraint that the path is self-avoiding. To find the dense folds we consider all configurations in simple confinements, such as those in an $l\times m\times n$ box (notice that our box size indicates the number

TABLE I. $2\times2\times2$ box, choice number $Z=4$ and $Z=5$. Number of configurations as a function of elements. N_{dense}: dense configurations with maximum number of nearest neighbors. N_{total}: total number of configurations. Compare with Tables II and III.

N_{elements}	N_{dense} $(Z=4)$	N_{total} $(Z=4)$	N_{dense} $(Z=5)$	N_{total} $(Z=5)$
1	1	1	1	1
2	1	1	1	1
3	1	4	1	6
4	6	15	8	26
5	9	53	12	104
6	8	161	8	372
7	6	444	6	1236
8	24	1100	36	3763
9	76	2590	164	10890
10	84	5560	192	28664
11	48	11412	146	72416
12	120	20384	584	162364
13	722	35280	3984	354036
14	988	52078	6488	674236
15	424	76116	3264	1264156
16	396	90936	5464	2036904
17	172	106728	4220	3267244
18	160	97362	8440	4399672
19	2908	87696	115084	5929000
20	6366	57460	313360	6452560
21	1752	36684	141188	7011716
22	3300	15088	496648	5731068
23	656	5812	316352	4606488
24	848	924	865544	2399816
25	0	0	780624	1128736
26	0	0	206692	206692

of elements, so our $2\times2\times2$ box is the same as a $3\times3\times3$-bead box). The dense chain configurations are easily derived from the total number of occupied nearest neighbor sites to any element's end point on the chain. All the configurations that fit into a box of a given size are counted. This gives a large number of folds, as can be seen on Fig. 3, for a $2\times2\times2$ box (thin lines for various box sizes) and shown explicitly in Table I. Next we find among those all that are densely packed in the sense of having the largest number of neighbors. This is plotted as the heavy full line. We notice it is very irregular with dips at numbers we shall call *elemental* ''magic'' numbers. Similar dips were found in a count for the filling of a 2D plane [22,39]. A simple analysis shows that the dips in 3D correspond to (in sequence) filling a $1\times1\times1$ box at the number of elements $N=7$, a $2\times1\times1$ box at $N=11$, a $2\times2\times1$ box at $N=17$, and $3\times$ packed $2\times1\times1$ boxes at $N=23$. It is not possible with $Z=4$ to completely fill a $2\times2\times2$ box. This can be done if we allow also straight continuation of the elements, i.e., using $Z=5$. The results for the dense folds $Z=4$ and $Z=5$ are shown in Fig. 4. The number of folds are much larger in the latter case. A scaling and mean field theory [50] of this problem gives the estimate that the number of folds for N elements increases as $(Z/e)^N$, where Z is the choice number, in our case $Z=4$, and $e=2.7183$. For a protein with nine secondary structures and consequently eight interconnecting

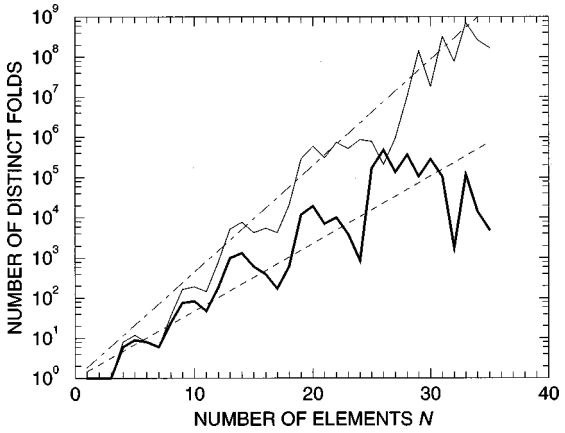

FIG. 4. The number of distinct dense folds for coordination number $Z=4$ in a $3\times2\times2$ box, fat line. Notice the deep minima at *magic* numbers at the closed configurations. Notice the deep minima at *elemental magic numbers* N_m: 7,11,17,23,31,35, etc. at the closed configurations; these correspond to *secondary magic number* of elements $N_s=(N_m+1)/2$: 4,6,9,12,16,18, etc. The dashed line represents the mean field estimate for $Z=4$: $(Z/e)^N$. The corresponding results for $Z=5$ are shown as the thin an the dashed-dotted lines.

loop elements we have $N=17$, and the above theoretical relation gives the number of folds as $(4/e)^{17}\sim711$. This is already a quite small number. However, the discreteness gives rise to *magic* numbers at which there are particularly few, different folds. Although the mean field theory represents the average data well, there are systematic deviations for large N. This is because we have not included closed folds in very elongated confinements, such as, e.g., a $4\times1\times1$ box, which we exclude since they are not ''globular,'' although they do fulfill the simple neighbor criterion. We have given the exhaustive count of the dense and the total number of configurations for box sizes up to $3\times2\times2$ for $Z=4$ in Table II, and for $Z=5$ in Table III. To obtain the exhaustive dense count for a certain number of elements a few numbers have to be added for less globular box sizes [51]

In the context of protein folding Dill [11] has analyzed and found the effective choice number for a typical protein to be $Z\leqslant3.8$, although this is an average for all residues. Based on this and the above argument [48], we find that the simple case we have described with $Z=4$ is in fact the most relevant for real proteins. From our numerical calculations we can then estimate how many distinct fold classes there are to be found. If we, for reasons given in the next paragraph, restrict ourselves to domain structures with $N\leqslant17$ we find in total 3906 possible, distinct globular fold classes. This is close to Chothia's estimate of 1000, based on the heuristic argument [24]. When increasing the number of elements in a domain beyond $N=17$ by just a few the number of possibilities increases dramatically. It is interesting that our estimate based on completely different arguments is close to Chothia's, and reduced from the astronomic numbers that would arise from the most direct enumeration [14]. The fact that we get a slightly higher value, if significant, could indicate that nature may not have used all possibilities available by structural ''symmetry'' during the course of the evolution. We may have to further impose a designability criterion

TABLE II. $3\times2\times2$ box, choice number $Z=4$. Number of configurations as a function of number of elements. N_{dense}: dense configurations with maximum number of nearest neighbors (for the exhaustive count the numbers in Table IV must be added). N_{total}: total number of configurations. N_{nn}: maximum number of nearest neighbors. The final column shows when particular simple subunits are maximally filled.

N_{elements}	N_{dense} $(Z=4)$	N_{total} $(Z=4)$	N_{nn}	Comments
1	1	1	2	
2	1	1	4	
3	1	4	8	
4	6	15	10	
5	9	57	14	
6	8	207	18	
7	6	731	24	$1\times1\times1$
8	24	2376	26	
9	76	7193	30	
10	84	20112	34	
11	48	53232	40	$2\times1\times1$
12	184	130872	42	
13	978	305074	46	
14	1312	655566	50	
15	602	1349200	56	
16	396	2519548	60	
17	172	4547644	66	$2\times2\times1$
18	616	73391244	68	
19	11782	11585834	72	
20	19354	16095254	76	
21	6972	22105158	82	
22	10016	26351888	86	
23	3902	31361586	92	
24	848	31658298	96	$2\times2\times2$[a]
25	166120	32057672	98	
26	478392	26652332	102	
27	134136	22350538	108	
28	365704	14585004	112	
29	105246	9643600	118	
30	283660	4535516	122	
31	102870	2185692	128	
32	1752	629544	134	
33	115808	195360	138	
34	14016	25460	144	
35	5006	5006	150	$3\times2\times2$
36	0	0		

[a]It is not possible to fill the $2\times2\times2$ box with the optimal 26 elements using $Z=4$.

TABLE III. $3\times2\times2$ box, choice number $Z=5$. Number of configurations as a function of number of elements. N_{dense}: dense configurations with maximum number of nearest neighbors (for the exhaustive count see [51]). N_{total}: total number of configurations. N_{nn}: maximum number of nearest neighbors. The final column shows when particular simple subunits are maximally filled.

N_{elements}	N_{dense} $(Z=5)$	N_{total} $(Z=5)$	N_{nn}	Comments
1	1	1	2	
2	1	1	4	
3	1	7	8	
4	8	30	10	
5	12	142	14	
6	8	632	18	
7	6	2645	24	$1\times1\times1$
8	36	10134	26	
9	164	36782	30	
10	192	124298	34	
11	146	401013	40	$2\times1\times1$
12	796	1203304	42	
13	5172	3460894	46	
14	7696	9150100	50	
15	4268	23413384	56	
16	5464	54574722	60	
17	4220	124465702	66	$2\times2\times1$
18	20528	256696224	68	
19	286044	523201896	72	
20	590112	956157616	76	
21	304504	1740791038	82	
22	740264	2808524872	86	
23	523094	4540269028	92	
24	865544	6395425216	96	
25	780624	9062517568	102	
26	206692	10917458588	108	$2\times2\times2$
27	936888	13261852260	110	
28	10182968	13192946730	114	
29	142150014	13246041324	118	
30	18009792	10243424132	124	
31	322585300	7986809176	128	
32	76012112	4275862868	134	
33	711760872	2291702688	138	
34	265368752	632026676	144	
35	169462384	169462384	150	$3\times2\times2$
36	0	0		

[43,52] or a functional criterion to reduce the number somewhat.

We further believe that there is a connection between the simple geometrical ''preferred'' numbers found in the close packings and (1) the breaking up into domains and (2) the preferred number of residues in protein domains. The local minimum at $N=17$, corresponding to nine secondary structures is relatively well pronounced and the next minimum is anomalous. There is also a well pronounced minimum at the magic number $N=35$. The $N=35$ structure is confined in a $3\times2\times2$ box. An analysis of the folds shows that a large part is formed of two folds of the $N=17$ domain interconnected by just a single element, i.e., $2\times17+1=35$. This explains why the domain formation is a natural consequence of the discrete packing problem and that the natural choice for a domain size contains 17 elements, which in turn implies a certain length in terms of residues.

There is experimental support for this, which has been seen by studying the statistics of the length distribution of protein chains [15] in the databases. Those distributions show optima in protein length around 125, 250 amino acids (aa), etc. for eukaryote and similarly 150 aa and 300 aa for prokaryote. The origin of this remarkable periodicity has yet to be explained in detail. It can have something to do with the topology of the polypeptide chain in early stages of pro-

TABLE IV. Additional dense configurations for $Z=4$ to be added to Table II for obtaining an exhaustive count. It is arising from filling the nonglobular boxes indicated.

N_{elements}	$4\times1\times1$	$4\times2\times1$	$3\times3\times1$	N_{nn}
19	508			72
25		50318	87558	98
26		83912	169136	102
27		34652	67498	108
28		40404	110468	112
29		19074	45086	118
30			67176	122
31			36430	128

tein folding [53,54] or the phenomena could be a remanence of the DNA-RNA structures. Here we propose that these periodic optima are related to the packing of the polypeptide chain at the later stages of protein folding. As to be demonstrated below, the position where the curve in Fig. 4 has a minimum is a special ''economical'' configuration for domain sizes. They are the most common protein domains, the length of which is given by the amount of residues in the secondary structure (α,β) elements and loops (of length of around 11-6-4 residues, respectively, Fig. 6), plus the beginning and end segments. This gives for a $N=17$ element domain the following number of residues: for a pure α domain $\sim$150 residues and for a pure β domain $\sim$100 residues. Based on the average size of the elements, the magic numbers therefore also rationalize why the size of the domains in terms of amino acid units [15] is as preferred by nature. It is interesting that this number is also in accord with the overall thermodynamic theory [11] for the effect of hydrophobic forces acting on a polymer chain.

One might argue that the restriction of the chain to have elements being only orthogonal to the preceding one is too limited in the sense that two consecutive parallel elements could also be considered and counted for in the total energy. To do that, we may include the following term to the previous Hamiltonian:

$$\mathcal{H}_{\text{straight}}=-\sum_{P} L_P \hat{\mathbf{e}}^P\cdot\hat{\mathbf{e}}^{P+1}-\sum_{p} \ell_p \hat{\mathbf{e}}^p\cdot\hat{\mathbf{e}}^{p+1}. \qquad (7)$$

We have carried out a study where we included the case with coupling constants L_P and ℓ_p. This means that when it is being decided whether an element is orthogonal to the previous element in the plane (J_P,j_p) or out of the plane (K_P,k_p) we also include the possibility of the element going straight ahead from the previous one. This extra move possibility gives rise to a new list of configurations shown in Table III. The possibility of including the straight moves $(Z=5)$ gives a much larger set of unique configurations. However, the behavior exposed in Figs. 3 and 4 of minima at 7,11,17,... number of elements is still maintained in these extended numerical calculations (as can be seen by comparing Fig. 3 and Fig. 4, for $Z=4$ the magic number dips gets more pronounced in the larger box).

We have performed a series of calculations for different sizes of lattices in order to see the variation in the number of different configurations for optimal packing densities for all the possible sizes of chains. The numerical calculations are performed as an exhaustive search for all the possible configurations. In the table we pay special attention to the minima occurring at specific chain length and with a maximum of neighboring occupied lattice sites that appear in specific lattices and reappear in the sublattices contained in the former lattice. For example, the results for the lattice $(3\times2\times2)$ contain all the minima encountered in the smaller sublattices such as $(1\times1\times1)$ and $(2\times1\times1)$. Furthermore one can make a study of the statistics of optimal packed configurations for specific chain length as a function of different lattice sizes. As expected, the number of configurations with the magic number of elements for the $1\times1\times1$ lattice will remain the same for all greater lattices.

A. Graphical representations of the protein folds

Basically the philosophy behind our representation of folds is that the 3D protein structures can be represented in a unique way by a 1D string of coupling constants $(\pm J,\pm K,\dots)$. That is a unique name written with an 8-letter alphabet (which we have demonstrated may be extended to sixteen or more letters, when including more distinguishing features; the minimum is four letters). It is independent of rotations and moderate distortions (twists) of the proteins. We have given the prescription for how that can be done once the protein is partitioned in secondary structure elements. Another protein with the same number of elements but with a different string representation will have the same energy with respect to the hydrophobic forces, but could differ with respect to the hinge coupling parameters.

The projection of the actually observed (twisted) structures to the high symmetry representation can be made by visual inspection of the stereographic pictures. However, for a more systematic approach we have constructed a computer program that can convert a set of protein coordinates in the PDB (Protein Data Bank) format into our representation of ordered chain elements on a regular lattice. The actual structures are *locally and consecutively* rectified to the rectangular representation. The representation can be given in a nice graphical form and yields a systematic name.

V. MAGIC NUMBERS

We now turn to the question of an atomistic grouping of packed structures of protein chains as considered in the previous section. From an analysis of packing and the effect of hydrophobic forces [12] we shall try to understand the appearance of ''magic numbers'' and test the paradigm by a statistical analysis of available structural data. Magic numbers are well known in graph theory and packing of hard spheres. For the 2D square lattice the occurrence and the origin of the magic numbers were discussed explicitly by Chan and Dill [22]. Later further studies were performed [39]. The studies of two-dimensional lattice animals give some guidelines for the statistical behavior of proteins, however, for a property such as magic number it is imperative to study the relevant 3D problem. We argue that the 3D magic numbers have a profound physical meaning for the proteins. The fact that the 2D elemental magic numbers [22] are quite different from the 3D ones actually corroborates our model, as will discussed below in Sec. III C. After this paper was

ELEMENTAL MAGIC NUMBERS
7, 11, 17, (24), 32, 35, ...
SECONDARY MAGIC NUMBERS
4, 6, 9, (12), 16, 18, ...

FIG. 5. The predicted magic numbers.

completed we were notified that 3D counts for $Z=5$ actually had been made earlier and used to analyze the bead model [23] for N up 13 and for $N=27$. The results agree in all details with ours.

The magic numbers found for the 3D lattice animals are not very sensitive to deviations from a linear weighting of the neighbor count, which is still consistent with the globular structures. The magic numbers in our model are *universal* in the sense that they do not depend on the specific, chemical interactions between the amino acids: neither between distant parts of the chain nor along the backbone—they are dictated by the hydrophobic, confining forces.

Figure 4 shows the exact, exhaustive enumeration of all possible dense folds for elements up to $N=35$ in a $3\times2\times2$ box. For $N=17$ there is a pronounced minimum with only $p(17)=172$ distinct and predictable folds. The mean field theory, giving 711, overestimates this grossly. Between the magic numbers the abundance is, on the other hand, much larger. The magic number at $N=7$, corresponding to the 4-helix bundle, is a close packing of a $1\times1\times1$ box, which we call an A box. The next closed confinement is the $2\times1\times1$ box, which we call a B box. Magic numbers at $N=11$, 17, 23, 32, and 35 can be understood as the optimal packing in closed polyhedra (analogous to shells) consisting of 1, 2, 3, 5, and 6 B boxes. The minimum at $N=24$ corresponding to a best filling of a 4 B box is anomalous because the $2\times2\times2$ box cannot be completely filled with the optimal 26 elements for $Z=4$. With this in mind the predicted elemental and secondary magic numbers are summarized in Fig. 5. For $Z=5$ the 4 B box can be packed with 26 elements. This would correspond to 13 secondary elements.

The magic number folds represent closed confinements having minimal surfaces and are thus energetically favorable from the point of view of the hydrophobic forces. They have a clear energy separation from other, neighboring folds. This is, according to the theory by Shakhnovich [40], a necessary condition for them to be able to fold rapidly (see also [41]). The configurational entropy for a fold at the magic number is low, and allows the large entropy of the extended chain to be exchanged by energy gain, without significant change in free energy. This indicates that proteins with the magic number of elements could be more stable and fast folding than others. In the following we are going to test this by comparing with experimental findings and by thermodynamic analysis.

A. Statistics of secondary structure abundance in nature

In order to be able to evaluate the relevance of the numerical calculations and compare the computer results with real protein data we need to perform the statistics of how many proteins occur with a certain number of secondary structure elements. In other words, we would like to see if there has been a selection pressure such that nature has a preference for building up proteins of a certain number of helices and strands.

An important part in getting reliable statistics of occurrences in genetics and in molecular biology is to get a database with no biases, e.g., a group of proteins not containing a particular amount of a secondary structure. It is therefore most appropriate to resort to data sets that have been selected especially for a nonbiased content. The data sets used for training and testing neural networks on secondary structure predictions are convenient since they constitute a standard reference for the whole molecular biology community.

We have used the standard set of 136 proteins with a sequence similarity below 25% selected from PDB by Rost and Sanders [55], originally used for secondary structure prediction. The secondary structure assignments are made by the DSSP algorithm [47] in which the hydrogen bond potentials (being the physical basis for the secondary structure stability) are calculated from 3D atomic coordinates. This results in assigning a particular type of secondary structure character to each residue in a given protein, indicating that the residue participates in that type of structure. Secondary structures of a given type are identified as such if they contain at least 4 consecutive residues. The decision of how many residues constitute a secondary structure is crucial for the statistical analysis of the abundance of secondary structures. In Fig. 6 we have displayed the size distribution of secondary structures for all known proteins in the complete PDB database. The helix distribution (a) has its maximum spread out over a plateau stretching from 4 to 12 residues, the β-strand distribution (b) has a maximum around 3-4 residues, and the loops (c) a maximum at 4 residues. This clearly gives support to defining secondary structures as containing at least 4 consecutive residues. We have also performed statistics with a definition of helices containing more than 4 residues as a minimum requirement, but that did not alter significantly the statistics of secondary structure abundance. In making the secondary structure statistics of Fig. 6 we have counted α-helix and 3_{10}-helix assignments as one type and all β strands as another and then counted them all together.

In Fig. 7 we have displayed the found abundance of the secondary structures as a function of their number on the basis of the Rost and Sanders database [55]. The curve clearly shows local maxima in the abundance that correspond to the optimal packing we find theoretically. We find optimal abundance at the following number of secondary structure elements: $N_s=4,6,9,16,18$, etc. The statistics for the higher values is probably not reliable. The numbers correspond to the number of elements being $N_m=2N_s-1=7,11,17,31,35,\ldots$. Notice the large coincidences with the *elemental magic numbers* obtained from our computer studies, Fig. 4. Only $N_s=12$ is missing, probably because of the small database used or because N_s is anomalous for $Z=4$ and $N_s=13$ for $Z=5$. The found optima are stable as to what size of the database we use; e.g., the first half of the data set has roughly the same distribution as the second half of the set. This means that protein folds with a magic number of elements, and a corresponding magic number of secondary structure elements are more abundant. That again means that the respective fold classes are larger, i.e., are containing more members.

B. Magic numbers and the Euler characteristics

How can we understand and construct the series of magic numbers for packing of the protein chain? As we have seen

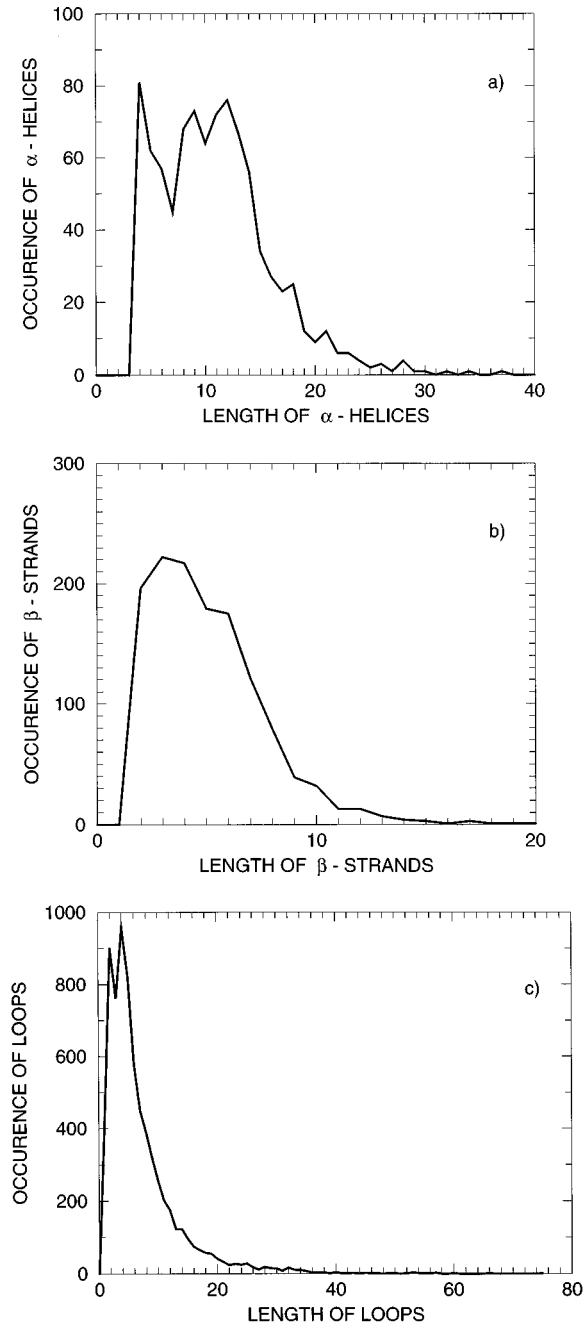

FIG. 6. (a) shows the statistics of the lengths of α helices, (b) that of β strands, (c) that for loops obtained by analyzing the complete PDB data base.

the magic numbers of secondary structure elements occur when the number of dense packings has a local minimum. At the position of a magic number there is a maximal jump in the total number of closest neighbors around each lattice site occupied by the chain. We shall argue that the magic number occurs when the chain forms a closed surface (box) within the lattice. A good example is the 4-helix bundle at the magic number, $N_s=4$, corresponding to $N_m=2N_s-1=7$ chain elements, which form a closed A box $(1\times1\times1)$ that can be embedded in any other larger lattice.

For closed surfaces we have the Euler equation that connects the number of corners c with that of edges e and faces f. The formula is

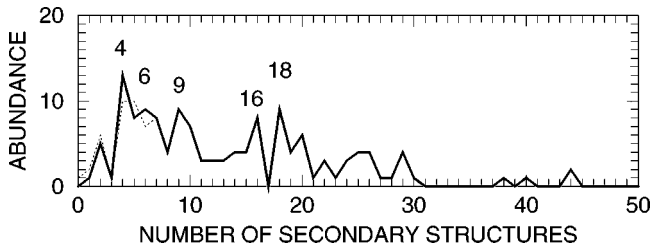

FIG. 7. Statistical abundance of proteins with $\mathcal{N}$ secondary structures [32].

$$\chi=c-e+f=2-2g, \tag{8}$$

where g is the genus number. We shall in the following only be considering surfaces with no genus $(g=0)$. In case the total surface of the chain configuration is not closed or the body has buried corners the equation is not fulfilled but becomes instead

$$\chi=c-e+f=m\neq2, \tag{9}$$

where m is any natural number.

One can get a clue on where magic numbers occur by calculating the density of chain elements through the total sum nn of the number of nearest neighbors that the end points of the elements (vertices) on the chain have. At the magic number the number m in the Euler equation is two, meaning that the chain configuration makes up a closed surface (or box), and the jump in the number of neighbors is optimal $\Delta nn=6$. The next magic number is obtained by adding a new closed box to the other in the lattice and see when it is filled out by the chain. For the case of the A box alone the *elemental magic number* can only be $N_m=7$, which is the number of vertices minus one. But in the case of the B box $(2\times1\times1)$, which contains the A box two times, we obtain the next magic number: $N_m=11$. In Table II we show the explicit number of configurations and the maximum number of neighbors.

Let us try to examine in detail the cases where the chain is configured around an A box and then has a few extra elements as shown in Fig. 8. As we saw the elementary box was filled out well by the 4-helix protein chain and satisfied the Euler condition with 8 corners, 12 edges, and 6 faces. With an extra element added to these chain configurations we obtain one more corner and one more edge but no extra faces. We can count the extra nearest neighbors as being simply the sum of all the attributes, $\Delta nn=+2$. With two more elements (see Fig. 8) we have 2 extra corners, 3 extra edges, and 1 extra face. By adding these extra quantities we get 6 minus the 2 from the last case, making the extra nearest neighbors $\Delta nn=+4$. By adding one more element we end having again $\Delta nn=+4$. If we sum up all the corners, edges, and faces for these cases with extra elements including the extra corners, etc. we cannot satisfy the Euler relation for this ex-

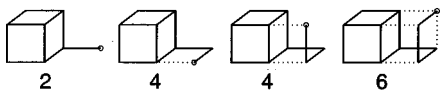

FIG. 8. This graph shows the increment in the number of neighbors when adding extra elements to a closed box configuration.

tended surface structure that is not closed in these cases as anticipated above. If we, however, add one more element [i.e., all together 4 elements on the ''magic'' $(1\times1\times1)$ box] we end up getting 6 more corners, 8 more edges, and 4 more faces, which altogether is $c=12$, $e=20$, $f=10$, which we see satisfies the Euler relation again. We have arrived at the next magic number configuration of 11 chain elements corresponding to 6 secondary structures. Furthermore we can count the extra nearest neighbors obtained by this configuration as being $\Delta nn=+6$, which is precisely what is observed in the Table II of the numerical calculation of chain configurations. Going to the magic number configurations the average number of nearest neighbors increases to $+6$ from the previous configurations with one element less. We have found a procedure for determining a magic number occurrence by using the Euler relation and counting the extra content of corners, edges, and faces, which thus gives us the number of nearest neighbors and hence the density of the chain configuration. We can extend this prescription to more complicated lattice boxes.

For up to 60 elements the magic numbers follow the filling behavior of the B box. That is because an A box can be placed several ways around a larger box (consisting of 1–12 B boxes), which increases the number of possible, different folds. Without having done the actual numerical enumeration of folds, we predict—from the fact that 8, 9, and 12 B boxes provide especially closed confinements with maximal number of neighbors per element and minimal number of faces per element—that the higher elemental magic numbers most likely include $N_m=44$, 47, and 59, corresponding to $N_s=22$, 24, and 30. However, as we saw for the 4 B box the problems with the actual folding of the chain into the confinements may alter the simple estimate somewhat. Further, the higher numbers may not be relevant for proteins in view of their tendency to form agglomerates of domains of smaller structures.

In conclusion, the magic number configurations satisfy the Euler relation, due to the minimalization of surface area compared to that of volume. This is due to the hydrophobic forces that tend to minimize the number of hydrophobic side chains on the surface of the chain configuration.

C. Magic numbers and other lattices

Although we have emphasized that the simple cubic lattice sc is the minimal description of the chain of elements in 3D space [48], it is of course not given that nature has restricted itself to that, and therefore it is of interest to see whether the obtained magic numbers are just specific to the discussed sc lattice description. The magic numbers arise because certain closed boxes are singled out as being particularly favorable with respect to the hydrophobic forces, as included by the procedure of counting the neighbor. For the simple cubic lattice a magic number configuration is particularly favorable relative to the unfolded states, since the nearest excited state costs 4 neighbor bonds, whereas for the nonmagic configuration the cost is only 2. That gives the important larger energy gap for the magic folds. A magic fold with N elements is also favored with respect to the nearby dense folds with $N-1$ and $N+1$ elements, as can be seen on Fig. 9. This shows that the number of neighbors per elements is consistently higher for the magic folds. The hydrophobic forces will thus tend to favor the formation of the magic number of elements in the molten globule stage, if the element number still fluctuates. Further the degeneracy $p(N)$ is smallest for the magic number of elements, since both $p(N-1)$ and $p(N+1)$ are larger in the number of ways one can distribute either a missing or an added element.

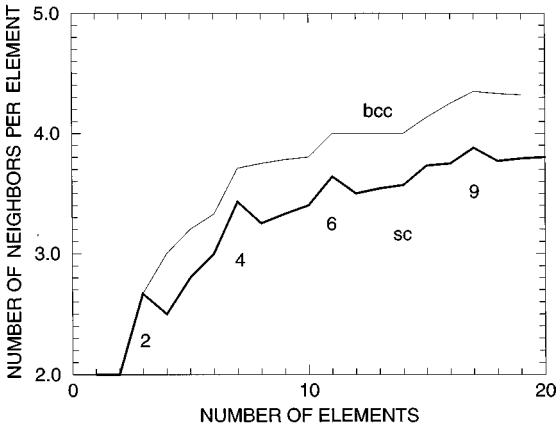

FIG. 9. The number of neighbors per element for various number of elements; fat line: for the simple cubic lattice (sc), thin line: same for the bcc lattice.

If we now consider a bcc structure, see Fig. 10, the difference between this and the sc structure is that the lattice ''unit'' cell is deformed relative to the sc structure and an extra nearest neighbor bond is formed (dotted line, along a body diagonal). The topological difference between a sc and a bcc structure is simply that for packing on a bcc lattice we in addition allow an element to be placed along *one* of the body diagonals in the sc cell. The choice number is (if we do not include going back or going straight) $Z=6$. Apart from the additional possibilities for placing an element, the same boxes are preferred as for the sc case, having the magic number of boxes favored by 4 neighbor bonds relative to an excited structure. Since the maximum number of elements in a box is given by the number of vertices N_v as $N_m=N_v-1$, the magic number of elements of the bcc lattice are identical to those for the sc lattice. The actual degeneracy number is larger: $p_{\mathrm{bcc}}>p_{\mathrm{sc}}$, due to the larger value of Z. Another difference is that for the bcc lattice we do not have as clear a stabilization relative to the neighboring number of elements, which can be seen on Fig. 10, thin line. This shows only shoulders instead of clear maxima in the number of neighbors per element.

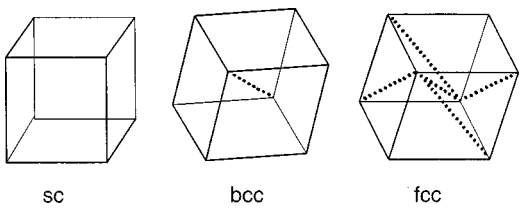

FIG. 10. Left, simple cubic lattice. Middle, the box squeezed so that an extra nearest neighbor bond appear. This is a representation of the bcc lattice. Right, the box further squeezed so five extra neighbor bonds appear. This is a representation of the fcc lattice.

For a closed packed structure, as for example the fcc, our treatment of the hydrophobic forces (as neighbor count) does not single out any preferred boxes. The number of neighbors per element increases monotonically as does the number of ways to distribute the elements i.e., for the fcc structure we do not find any magic numbers. However, a closed packed structure is contrary to our picture of the parent phase, and packing on a closed packed structure is therefore not relevant in the present context. Murzin and Finkelstein did consider the packing on closed packed polyhedra [35], which is consistent with their attempt to describe the twisted state, not a parent state. As shown in Fig. 10 the topological difference between a sc and a fcc packing is simply that in the fcc we allow in addition an element to be placed on *one* body diagonal and on *four* face diagonals, and the choice number is $Z=10$. It is now easy to generalize to other equal length lattices. Only those that are not closed packed are of interest for forming the parent phase (with respect to the neighbor count criterion). For example, for a simple hexagonal crystal the relation to the sc is that in addition *two* opposite face diagonals are allowed for the placing of the elements, and $Z=6$. It seems that some observed structures are most naturally described if we occasionally allow elements, in particular loops, to be placed on diagonals. A recent example is the ''normal'' form of the scrapie prion protein. This, according to the model by Huang *et al.*, is a nicely twisted 4α-helix bundle, which is different from those described in Sec. III B for the sc projection [56].

Although it goes beyond the scope of the present paper, a remark about relaxing the equal length assumption is in order here. For α helices and β sheets an almost parallel packing with small connecting loops is often found; see, e.g., Fig. 1 in Ref. [32] or Fig. 2 in this paper. A packing of a small number of secondary elements on a tetragonal lattice then seems to be a more natural choice for a parent state (it leads to a reduced model also used for the Martensitic transformation [57]). It is interesting that the magic numbers N_s for the secondary structures remain the same (see a detailed discussion in [58]), as long as the simplification makes sense, and that they are robust with respect to how the loops are placed and how long they are.

We conclude that for the packing on lattices for which our description of the hydrophobic forces in terms of neighbor counting singles out closed boxes, the magic number remains those we have discussed for the simple cubic lattice. However, the total number of possible structures selected only on the discussed basis will be higher.

VI. THERMODYNAMIC THEORY FOR PROTEIN FOLDING

So far we have only used the Hamiltonian Eq. (4) for enumerating the distinct folds found according to the hydrophobicity criteria, without explicitly writing down a Hamiltonian for the hydrophobic forces. This and the Hamiltonian for the short ranged (twist) forces will be discussed in this section. First we emphazise that it is likely the protein folding problem is an essential nonequilibrium phenomenon in the thermodynamic sense, and an energy function is only describing part of the process. Since the dynamically unknown time interval is large, ranging from 10^{-10} to 10^{-3} s (from either molecular dynamics calculations or experiments) a proper theory for the dynamic folding processes in that interval is still far fetched. At most one can make a scenario, the details of which are to be resolved experimentally or computationally. First, the temperature is not a well defined concept, and can be replaced by the properties of the solvent; even at room temperature one can fold and unfold proteins by varying the amount of denaturants. However, with this in mind, let us follow common practice [5,7] and use the word temperature as indicating a measure for the degree of folding. We envisage the following scenario in line with recent observations [3,4,20,21].

At high temperatures the protein will be in an *extended* state because of the large phase space for this. When cooling down, the protein will start to form the α helices because there is a clear chemical energy gain by forming hydrogen bonds between every third amino acid, and also a certain hydrophobic gain because of the contraction.

A. The molten globule and the parent states

The gross partitioning of the chain in secondary and intermediate structural elements is completed at the next stage [4]. Any interaction between the elements is supposed to be switched off by screening effects of the solvent. It is at this *Molten globule* stage we introduce our Hamiltonian Eq. (4) containing the hinge forces. They are of course generally very weak relative to other forces. How can they matter in the folding process? To demonstrate this, it is instructive to look at the Heisenberg magnet with the Hamiltonian:

$$\mathcal{H}_{\text{Heisenberg}}=-\mathcal{J}\frac{1}{2}\sum_{\langle ij\rangle}\mathcal{S}_i\cdot\mathcal{S}_j-h\sum_i\mathcal{S}_i^z, \tag{10}$$

where the sum is over all nearest neighbor pairs only. The strong interactions $\mathcal{J}$ cannot determine the spin direction in the fully rotationally invariant ordered state given by the dominant first term—but by introducing an infinitesimal field h in the z direction the rotational symmetry is broken. It is the weak global force that determines the overall structure; the strong force determines the details. This analogy is in fact deeply related to the present problem.

To see this we write the hydrophobic Hamiltonian as

$$\mathcal{H}_{\text{hydrophobic}}=-V\frac{1}{2}{\sum_{\langle ij\rangle}}'\sigma_i\sigma_j-\mu\sum_i\sigma_i, \tag{11}$$

where the sum is over all nearest neighbor pairs on a (large) cubic lattice and V is the (essentially hydrophobic and hydrogen bonding) energy gain for forming a secondary structure or loop. This appears to be a regular Ising model in terms of the occupation variables $\sigma_i=0$ for unoccupied sites and $\sigma_i=1$ for occupied sites. These are describing the $N+1$ vertices or hinge points of a protein consisting of N elements. In fact the sites can also be just the location of the residue at which the hinge is going to be, even in the totally unfolded state. By the chemical potential we may control the fixed occupancy to $\Sigma_i\sigma_i=N+1$. The only new feature, indicated by the prime, is that the probability for finding a state with differently distributed occupied sites has to be augmented by the number of ways the points can be connected

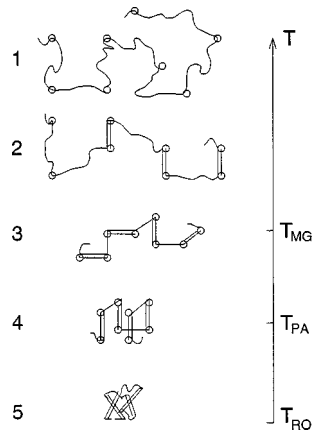

FIG. 11. A sketch of the five stage folding scenario from (1) the *extended state* at high temperatures to (2) a partly secondary structure forming stage, (3) the *molten globule* (MG) stage, (4) the *parent stage* (PA), and finally the *native*, twisted state at about room temperature T_{RO}. The double lines indicate formed secondary structures and single straight lines interconnections formed by loops. The ○ indicates the considered eight ''hinge'' residue positions.

by a single, self-avoiding line. This Hamiltonian describes schematically the above scenario for protein folding, as shown on Fig. 11 (for $N=7$ with 8 vertices, open circles).

(1) At $T\sim\infty$: A fully extended state where the potential hinge residues are sparsely distributed and no neighbor pairs are formed; thus no gain in hydrophobic energy.

(2) At $T_{\mathrm{MG}}<T<\infty$: The formation of secondary structures in the form of sparsely distributed neighbor pairs. A gain in hydrophobic energy of V for each formed nearest neighbor pair, representing a secondary or loop element.

(3) At $T=T_{\mathrm{MG}}$: The *molten globule* (MG), which is here defined precisely as the stage at which a chain is formed with all connected pairs. The energy gain is at least NV. It is still an extended chain of secondary structures and loops, having a mean (square) radius of gyration, according to polymer scaling theory [59] of $r_G=\langle R_g^2\rangle^{1/2}=aA_g^{1/2}N^{3/5}$, where a is the average length of an element, and $A_g(\sim\frac{1}{6}$ for sc) is the amplitude. This is nonuniversal. A universal ratio with the mean square end-to-end distance amplitude A_e is given by $A_g/A_e=0.1599$ (for a comprehensive overview over the statistics of self-avoiding random walks, see [60]). Using this and $\langle R_e^2\rangle$ and the estimate from [61] we find $r_G\sim\frac{2}{5}[(Z-1)/Z]\,aN^{1/2}$. The characteristic size scale is accordingly typically 20–30% larger, and the volume is a few times larger than that of the closed packed state.

(4) At $T=T_{\mathrm{PA}}$: The precise definition of the *parent phase* (PA) at which the chain with N elements forms a densely packed structure, in our chosen minimal boxes. This state then has $r_G\approx\frac{1}{2}Ra$, where R is the side length of the box depending on the number of elements (i.e., typically $R=1$ to 2 for $N=7$ to 25). The volume is only a couple of percent larger than that of the closed packed state. In the previous sections we have numerically calculated how many ways the points in such a box can be interconnected by a self-avoiding chain of N elements.

(5) Later we shall describe the Hamiltonian for the transition to the final, closed packed—so-called *native state*—taking place above or around room temperature T_{RO}.

At stage (3) to (4) each occupied site is a member of two pairs and it is meaningful to assign the hinge spin variable to the site, and thus introduce our hinge Hamiltonian, just for the occupied sites. The Hamiltonian Eq. (4) is a discrete one of the ''Ising'' type, where the spins can assume up to six different directions. Further, it is now describing a small ''cluster'' of only $N-1$ spins (when neglecting the outermost ones). Therefore, there will be no phase transitions in the true thermodynamic sense but rather smooth transitions from one state—or rather stage—to the other. For simplicity in discussion we map the potential native fold (one of the densely packed states) onto a ferromagnetic Ising chain with spin variables I_n. As we have seen, with respect to the dominating hydrophobic forces this state is degenerate with a large number $p=p(N)$ of other states, which may be thought of as p different staggered Ising states for a protein with N elements. The lowest energy excitation for the chain, with respect to Eq. (4), is a soliton mode in which all spins to the left of one are flipped. This violates the value of only one letter in the chain (change in sign or type), whereas a single spin flip requires change of two bonds. To evaluate the hydrophobic energy cost of these excitations (and check that the chain is still self-avoiding) we construct the site occupancy on the basis of the spin sequence Eq. (1) and calculate the energy from Eq. (11). The degenerate models all share the high energy excitation phase space, the molten globule. However, the low-lying excited states are very different—in particular because a large number of excitations are prohibited by the nonoverlap constraint for the folds, and the energies of extended folds are augmented by hydrophobic energy. At moderate temperatures the states are essentially independent and separated by large energy barriers. We introduce this regime as a new intermediate stage. It is a volatile, high symmetry *parent* stage corresponding to the bcc phase. We suppose that the energy cost in violating the dense packing W (which is of the order of V) is much larger than any of the hinge forces. However, a given set of hinge forces (which may include the effect of chaperones) sum up to give maximum energy gain for (most likely) the potential native fold. The $p-1$ other states will have a higher energy according to how many letters in the name have been violated. The effect is like that of the uniform field h in Eq. (10), and it is not sensitive to whether the hinge forces fit exactly to the final fold. So, without frustration the p-fold hydrophobic symmetry is broken. This demonstrates a natural relation between the sequence information and a preferred folding into the high symmetry fold corresponding to the native one [62].

B. Transition to the native state

At lower temperatures the folding process proceeds towards the experimentally observed *twisted*, so-called native structure. Only at this stage is the water supposed to diffuse out and leave a problem for the optimization of the short range chemical forces between neighboring elements. That is the problem addressed by Murzin and Finkelstein [35]. To describe this in our model we need an extra term in the Hamiltonian, just as for the Martensitic problem. Let, as described above, the parent state be represented by a p-fold degenerate effective Ising model with interaction parameter W:

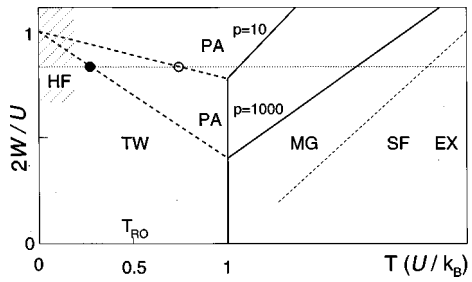

FIG. 12. A sketch of the phase diagram for the protein folding. Full lines represent continuous transitions (for $N\rightarrow\infty$), while dashed lines are discontinuous transitions. Two cases are shown one for large values of $p>2$, and one with very large values of $p\gg 2$. The entropy contribution depresses the phase separation line between the PA and the TW phase, at most for the large p. For a fixed ration $2W/U\sim 1$ there is a transition between the TW, the PA, the MG and the extended phases. The hatched region, marked HF, indicates structures determined by the hinge forces.

$$\mathcal{H}_{\text{parent}}=-\frac{1}{2}W\sum_{\langle n,n'\rangle} I_n I_{n'}, \tag{12}$$

where I_n are the reduced hinge spin variables (giving the changes relative to one of the considered p parent ground state configuration), and the sum is over all sites in the chain. This schematically represents the hydrophobic forces including the chain constraint and W (for water) represents the energy of the excited states mainly against the hydrophobic force. Therefore, W is of the order of a few times V plus the contribution from the hinge forces.

The native state is a twisted structure of one of the p particular states. Suppose it can only be twisted in very few equivalent ways, say 2. Then the native state can be represented by a normal transverse Ising model with degeneracy 2 and interaction constant U representing the previously neglected strong short range forces of chemical nature (disulfide bonds, etc.).

$$\mathcal{H}_{\text{twist}}=-\frac{1}{2}U\sum_{\langle P,P'\rangle} s_P s_{P'}, \tag{13}$$

where (A) the variables s_P could be occupancy variables as in the Ising model, with values $s_P=1$ or 0 according to whether two *secondary* elements are parallel, nearest neighbors yielding an energy gain U, or not (yielding no energy gain); the sum is only over the secondary elements. In a more realistic model for the twist (B) we could allow continuous variations in the variables and use $\hat{\mathbf{e}}^P\cdot\hat{\mathbf{e}}^{P'}$ instead of $s_P s_{P'}$, which could in turn allow for an energy also by twisting perpendicular elements (and possibly even in addition represent a slight move in space), yielding a Heisenberg type model. Presumably, that elaboration will not qualitatively change the results.

To understand the nature of the ''phase transition'' of the native folding processes we have to take a closer look at the entropy properties of the system. A very similar model was introduced and analyzed for the Martensitic problem [57]. It was recently simplified to two competing Ising models and further to a so-called degenerate Blume-Emery-Griffiths (DEG-BEG) model [63]. The latter was analyzed using mean field theory and Monte Carlo simulations. Values of p up to 6 were used, since in the Martensitic problem it is hard to imagine higher values. We here generalize the results, namely, to considering the case of a competition between a p-times degenerate Ising model (with a weak field), and where p can be very large, up to several hundred describing the fold degeneracy of the parent phase—and a transverse Ising model [the above case (A)] describing the twist of one of those phases. The only difference between the phase diagrams for the DEG-BEG model and the competing Ising models is according to the results of Refs. [57, 64] that in the latter case, there is a phase transition both between the W-stabilized (highly degenerate), as well as between the U-stabilized phase, and a disordered phase, which in our case corresponds to the molten globule phase. The entropy of the p-fold degenerate phase is according to [63] stabilized by a term $-k_B T\ \ln(p)$ in the free energy, with respect to both the disordered phase and the more ordered, twisted phase [65]. The free energies per site in a mean field approximation are

$$\begin{aligned} F_{\text{parent}}&=-\frac{1}{2}WM^2-k_BT\ \ln(p)/N+k_BT\left[\left(\frac{1-M}{2}\right)\ln\left(\frac{1-M}{2}\right)\right.\\ &\quad\left.+\left(\frac{1+M}{2}\right)\ln\left(\frac{1+M}{2}\right)\right],\\ F_{\text{twist}}&=\frac{N-1}{2N}\left\{-\frac{1}{2}Um^2+k_BT[m\ \ln(m)\right.\\ &\quad\left.+(1-m)\ln(1-m)]\right\}, \end{aligned} \tag{14}$$

where $M=\langle I_n\rangle$ and $m=\langle s_n\rangle$ and the prefactor in the last term is because we only include interactions between the secondary elements. Because we have mapped (approximately) the folding problem onto a known problem in statistical physics [57,63,64], we can without repeating the details of the derivation draw the schematic phase diagram for the protein fold (Fig. 12); the transitions across a dashed line are discontinuous (all or none). Results of using our Hamiltonian in Monte Carlo simulations and an analysis of the dynamical folding process after a quench from high to low temperatures are planned to be published elsewhere.

Depending on the relative strength of the various forces we then have different scenarios.

(1) If the short range forces are not sufficiently strong to force the energy barrier between the parent state preferred by the hinge forces and another of the p parent states, $2W/U>1$, the hinge force selected state will just be optimally twisted, but highly frustrated and not optimal from the point of view of the short ranged forces. This will be a state arrived at in a nonfrustrated manner, yet it will not be a state of minimal frustration, and not be in the lowest possible energy state. This situation is indicated by the hatched region, marked HF, in Fig. 12.

(2) If the short range forces are very strong, they can select the optimal one of the p available dense folds and overrule the hinge forces, corresponding to $2W/U<1$. Then there is a transition from the parent stage to a twisted state close to one of the p states accounted for by our theory. The

native structure is then given by the detailed interactions between the secondary elements. This will again neither be a minimum frustration nor in a minimum energy state because the major structuring was done by the hydrophobic forces.

(3) If $2W/U \sim 1$ there will be a competition between the two mechanisms. It could of course happen that the mechanisms during the course of evolution were selected so that both prefer the same state—without a competition. A competition would slow down the folding rate considerably. The insensitiveness to even quite substantial mutations [3,4] (replacements of parts of a sequence) could indicate that there is not at least strong competition.

(4) If the short range forces are very strong indeed, so strong that they can break up the already formed secondary structures, $2W/U \ll 1$, our analysis is less relevant, since the secondary structure count at the parent stage level may get seriously distorted. This limit is that which may be better described by the bead model. The native state for this case may be one of minimal frustration for all forces, but exceedingly difficult to find.

C. Preferred abundance of magic number proteins

We have above discussed the last two transitions molten globule↔parent↔twisted stages in general. Let us here consider the influence of the degeneracy factor p. We remark that for small p, i.e., the magic folds, the magnitude of U can be smaller (W/U larger) than for the other folds, and still cause a transition to the native phase for $T \geqslant T_{\mathrm{RO}}$. The value of W is given by the hydrophobic forces and should be relatively weakly dependent on the specific sequence constituting the involved elements, since the number of residues in each element is quite large (~ 10). On the other hand the value of U represents the total effect of the frustrated short range forces [divided by the number of secondary elements: $(N-1)/2$] between the various parts of the protein in its twisted, native state. If the magnitude of U can be small and still sufficient for ordering at room temperature, it indicates that the ordering into the native fold is not highly sensitive to finding an optimal solution of the frustration problem of matching neighboring sequence segments. Many different sequences can therefore do the job. Contrary to the Martensitic problem the interaction forces between the elements are highly frustrated and the energy gain therefore limited. We suggest that the elements are predominantly positioned by the hydrophobic forces with little chance for major rearrangements in the cases (1) to (3) discussed above. This would render a state susceptible to only ''local minimum frustration'' in terms of the theory discussed by Wolynes *et al.* [7].

In other words, we argue that for large p the transition between the parent phase and the twisted phase (native) will be depressed in temperature. Then it will require specially favorable constitutions (i.e., sequences of amino acids) of the elements to minimize the frustration in their mutual interaction, which is needed to stabilize the final twist order above room temperature. On the other hand, for the ''magic'' folds the restriction is much less severe because here p is relatively small, thus the constitution of the secondary elements is less critical and we would expect to have many more proteins belonging to the magic families. In a search for protein structures there is hence a greater chance to find them among the magic structures than among the few exceptional other ones. This explains the high abundance of the proteins with ''magic'' number of (secondary) elements. The arguments for the preferred abundance are in line with those given by Finkelstein *et al.* [52]. But they focus on the ''designability'' or ''multitude,'' and show that if a given fold can be made of many different sequences M_p the abundance is higher, since a large M_p reduces the ''free energy'' for such a fold by an entropylike term $-T^*\ln(M_p)$, where T^* is a ''configurational temperature''; unfortunately neither M_p, nor T^* can be calculated beforehand. That effect may be added (giving some further sorting) to the presently discussed entropy effect, which is arising from the degeneracy in packing of elements, and which as demonstrated is calculable. Our more elaborate arguments differ from those of Finkelstein *et al.* in particular with respect to the introduced phase transitions, and in that the temperature in our case is the real temperature (or reflecting a change in the solvent).

VII. DISCUSSION

A major asset of our theory is that all involved interactions are average quantities and therefore not crucially depending on specific realizations of sequences. It gives a basis for the classification and for the robustness against mutations. Further, it rationalizes the paradox that the direct forces taken one by one are strong, but the effect is small (because of cancellation between oppositely acting forces, the frustration). Our hydrophobic energy V is the average gain for forming a secondary structure involving of the order of ten residues, not for forming individual hydrogen bonds; Tanford has discussed the difficulties in evaluating the energy cost at that level [12]. It is of course an oversimplification to assume the same gain when assembling the elements, but it should be of the same order of magnitude. The interesting hydrophobic force W is even a further average of V. For our ''hinge forces,'' again, only the sum (or average) is of importance. Given that the secondary structures cannot be broken up totally at the twist stage, also only an average over the short range interactions resulting in U is of interest. It is clearly difficult to evaluate the effective interactions from first principles. However, the fact that the folding happens around room temperature T_{RO}, tells us that the energy scale of the parameters must be of the order of $\mathcal{E}_{\mathrm{RO}} = k_B T_{\mathrm{RO}}$, where k_B is the Boltzmann constant. This is equivalent to 0.60 kcal/mole (at $T = 300$ K). Suppose then that $W \approx 1\mathcal{E}_{\mathrm{RO}}$ and $U \approx 2\mathcal{E}_{\mathrm{RO}}$ (because it may be slightly stronger). For a given protein with N elements, the internal energies should scale roughly as N times these constants. We can now evaluate if the configurational entropy we have discussed resulting from the degeneracy p of the parent state can be of any significance. At room temperature the free energy contribution from this entropy is $\Delta\mathcal{E}_S(N) = -k_B T_{\mathrm{RO}} \ln[p(N)]$. For N ranging from 7 to 25, p ranges from form 10 to 1000. This gives an entropy contribution of from $\Delta\mathcal{E}_S(N \sim 7) \approx 2\mathcal{E}_{\mathrm{RO}}$ to $\Delta\mathcal{E}_S(N \sim 25) \approx 7\mathcal{E}_{\mathrm{RO}}$. In addition there are the variations according to the magic dips in $p(N)$. We find that the entropy per element is $\sim 0.3k_B$ for the magic number folds and $\sim 0.5k_B$ for all others; see Fig. 13. The discussed entropy thus gives an energy contribution of $\sim 30\%$, which is of

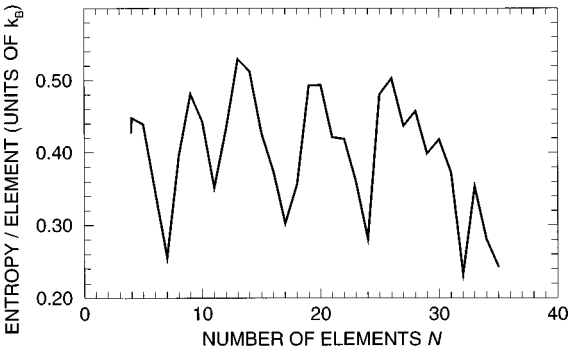

FIG. 13. The calculated entropy per element arising form the degeneracy of densely packed structures with respect to the hydrophobic forces. Notice the dips at the magic number of elements.

reasonable order of magnitude, and it is sufficient for causing a significant influence on the phase diagram.

Our models are of course extremely simplified. A major objection might be that one cannot strictly substructure the problem in the five stages we have assumed (which on the other hand seems to be in agreement with a considerable amount of experimental findings according to Jaenicke's conclusion [3,4]). However, we have demonstrated that within our model it is possible to have several (three) scenarios, simply depending on the ratio of the interaction constants $2W/U$. Of these we believe that the case $2W/U \sim 1$ most likely is the one prefered by nature, as it happens in the analogous Martensitic problem. That would give the most diversified transition scheme with the full sequence *native↔parent↔molten globule↔extended states* (of which we have not discussed the latter in detail). We have argued that the folding problem is a cluster (i.e., a small N) problem with no sharp transitions. In a recent study of magnetic relaxation in small Ising clusters [66] it was found that the transition from one state to another occurred by a nucleation mechanism, where the relaxation time is depending on the probability of forming a critical size droplet of the alternative order. A similar behavior is expected for the present models, and it is then in accord with the observations that the folding appears to happen in a concerted manner [21], with folding happening at several stages simultaneously around a first forming nucleus. However, our models will not be able to account for a scenario in which the folding occurs from the extended state directly to the native one only directed by the short ranged forces; this is handled by the bead model.

Finally, let us comment on the terminology problem of the folding intermediates. The experimental identification of a ''*compact globule with nativelike secondary structure and with slowly fluctuating tertiary structure*'' was probably first mentioned by Dolgikh *et al.* [10]. The presence of such a state, nearly as compact as the native state, is now established beyond doubt [2]. It corresponds well to our concept of the parent phase. Our state must be fluctuating sufficiently to experience the entropy in the p possible states, which are equally densely packed from the hydrophobic point of view. In the literature several names have been in use for such an intermediate state. In particular Ptitsyn has discussed this phase, see, e.g., [2] p. 265, and calls it a ''native-like molten globule.'' The concept of the p-times degeneracy of that state is not an element of the Ptitsyn model [2] nor of the later theory for side chain melting [67], thereby they differ significantly from our parent state concept, although they are supposed to be covering the same experimental regime. At stronger denaturation Ptitsyn proposed the term ''disordered molten globule,'' which would probably then be equivalent to what we have simply termed the molten globule, with a volume about three times the native [68]. Other names and concepts have been in use, such as folding intermediates and compact denaturated states [69] and others; see, e.g., [2]. However, none of the previous models has included the structural degeneracy, which in our theory leads to the magic numbers. Most of the experimental evidence for such states are indirect with respect to the actual structure.

It may not be easy experimentally to structurally assess if the parent structures are stable at higher temperatures, because the formation of the secondary structures may tend to break up, although in some cases an even higher content is suggested [2], p. 249. At the molten globule stage there may be proteins with an unstable number of secondary structures ''decaying'' into the stable ones, in quite close analogy to the shell model for nuclear matter.

VIII. CONCLUDING REMARKS

The hydrophobic forces cannot define a particular fold whereas the weak hinge forces set up a global force that will make a given protein fold predominantly in the right direction. We believe that the proposed Hamiltonian(s) makes sense in modeling the actual folding process from a certain stage. In our model we have at first neglected any forces between the secondary elements. This is an important conceptual aspect in our model for the not too late stages of the folding process. If specific amino acids on different elements could bind strongly it would fix the fold in any arbitrary configuration (imagine trying to fold double-glue-sided tape to a specific configuration). The physical justification for switching off these forces is that they could be screened by the water, which accordingly must have an important ''lubricating'' role to play during the folding. Only in the final approach to the dense fold is the water supposed to diffuse out and leave a problem for the final optimization of the short range chemical forces between neighboring elements. The result of that is undoubtedly the observed twistings and deformations of the actually observed structures. At that stage we have argued that the protein cannot make any significant refoldings, so most of these forces would be frustrated if they do not happen to match according to the underlying sequence. We thus have argued that a match is not instrumental in the folding process, whereby our model is very different from previous theories, which precisely focus on this problem of frustrating forces, and led to a comparison between the folding problem and the spin glass problem [7]. In our model there is no frustration in setting up the main part of the folding. The end result will necessarily be frustrated and therefore the native state is not the ground state for the chemical forces from an equilibrium thermodynamical point of view. It is interesting that there seems to exist a class of physics problems in complex systems in which ''partial ordering,'' of which we have discussed a particular case, is an important concept, which can be formulated mathematically [70], in more general terms. We emphasize that the

dynamical interpretation of the present model, which is susceptible to future experimental tests, is independent of the already experimentally supported structural classification of the native states, discussed in the main part of this paper. It would be highly interesting with more experimental information about the structure in the predicted *parent* stage.

ACKNOWLEDGMENTS

H.B. thanks P. G. Wolynes for enlightening discussions. Furthermore we would like to thank K. Rapacki, H. H. Stærfelt, and K. Lichtenberg for assistance with the numerical computations.

[1] L. Holm and C. Sanders, Science **273**, 595 (1996).
[2] *Protein Folding*, edited by T. E. Creighton (W. H. Freeman and Co., New York, 1992).
[3] R. Jaenicke, Prog. Biophys. Mol. Biol. **49**, 117 (1987).
[4] R. Jaenicke, *Current Topics in Cellular Regulation*, edited by E. R. Stadtman and P. B. Chock (Academic, New York, 1996), p. 209.
[5] A. V. Finkelstein and O. B. Ptitsyn, Prog. Biophys. Mol. Biol. **50**, 171 (1987).
[6] O. B. Ptitsyn, V. E. Bychkova, and V. N. Uversky, Philos. Trans. R. Soc. London, Ser. A **348**, 35 (1995).
[7] P. G. Wolynes, *Protein Folds*, edited by H. Bohr and S. Brunak (CRC Press, New York, 1995), pp. 3–17.
[8] L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. USA **37**, 729 (1951); L. Pauling, R. B. Corey, and H. R. Branson, *ibid.* **27**, 205 (1951).
[9] S. J. Prestrelski, A. L. E. Williams, Jr., and M. N. Liebman, Proteins **14**, 430 (1992).
[10] D. A. Dolgikh, R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G. V. Semisotnov, S. Yu. Venyaminov, and O. B. Ptitsyn, FEBS Lett. **136**, 311 (1981).
[11] K. A. Dill, Biochemistry **24**, 1501 (1985).
[12] C. Tanford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes* (J. Wiley & Sons, New York, 1980).
[13] J. Israelachvili and H. Wennerström, Nature (London) **379**, 219 (1996).
[14] C. J. Levinthal, Chem. Phys. **65**, 99 (1968).
[15] A. L. Berman, E. Kolker, and E. N. Trifonov, Proc. Natl. Acad. Sci. USA **91**, 4044 (1994).
[16] The data are available from the Brookhaven Protein Data Bank at www.pdb.bnl.gov.
[17] Information is available in the Swiss-Prot and the Trembl data bases at http://expasy.hcuge.ch/ and ftp://embl-ebi.ac.uk/pub/databases/trembl/.
[18] P.-A. Lindgård, J. Phys. IV (Paris) Colloq. **C4**, 3 (1991).
[19] R. J. Gooding and J. Krumhansl, Phys. Rev. B **39**, 3047 (1989); **38**, 1695 (1988).
[20] C. Redfield, R. A. G. Smith, and C. M. Dobson, Nature Struc. Biol. **1**, 23 (1994).
[21] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[22] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[23] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[24] C. Chothia, Nature (London) **357**, 543 (1992); J. Mol. Biol. **193**, 775 (1987).
[25] S. T. Rao and M. G. Rossmann, J. Mol. Biol. **76**, 241 (1973).
[26] J. S. Richardson, Adv. Protein Chem. **34**, 167 (1981).
[27] D. J. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).
[28] T. L. Blundell and M. S. Johnson, Protein Science **2**, 877 (1993).
[29] S. Pascarella and P. Argos, Protein Eng. **5**, 121 (1992).
[30] C. Sander and L. Holm, J. Mol. Biol. **225**, 93 (1992); **225**, 121 (1992).
[31] C. von Linné, *Fundamenta Botanica* (Linnean Society, London, 1736); *Species Plantarum* (Linnean Society, London, 1753).
[32] P.-A. Lindgård and H. Bohr, Phys. Rev. Lett. **77**, 779 (1996).
[33] An elegant discussion of the symmetry of proteins was written by P. G. Wolynes (unpublished).
[34] G. M. Crippen, in *Protein Folds* (Ref. [7]), p. 189.
[35] A. G. Murzin and A. V. Finkelstein, J. Mol. Biol. **204**, 749 (1988).
[36] A. V. Finkelstein and B. A. Reva, Nature (London) **351**, 497 (1991).
[37] J. B. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[38] M. Sasai and P. G. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[39] C. J. Camacho and D. Thirumalai, Phys. Rev. Lett. **7**, 2505 (1993).
[40] E. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[41] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4818 (1992); **89**, 9029 (1992).
[42] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[43] H. Li, R. Helling, C. Tang, and N. Wingren, Science **273**, 666 (1996).
[44] M Cieplak, S. Visveshwara, and J. Banavar, Phys. Rev. Lett. **77**, 3681 (1996).
[45] A. M. Lesk, *Protein Architecture* (Oxford University Press, Oxford, 1991).
[46] Nature only uses α-helices with one handedness, therefore we only assign one element to an α helix.
[47] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[48] For a rough navigation on a surface without a specified direction (north) four direction specifications are usually sufficient for short trips—with no angles given. A larger number of choices would be confusing and complicating if no lattice is given on which to move. If the length is given by the context the absolute minimal number of just two directions is sufficient (right and left). With our Hamiltonian Eq. (4) we have generalized this to representing rough navigation in a 3D space without specified directions (vertical and north). Therefore we believe that the choice of the cubic lattice is the minimal and natural choice for describing the highly flexible proteins.
[49] For the β-strand constants we could for example use instead of I,I,H,H the phonetic replacements $h_{\mathrm{ome}}, g_{\mathrm{o}}, c_{\mathrm{orrect}}, w_{\mathrm{rong}}$, and for L,L,ℓ,ℓ the replacements $m_{\mathrm{ore}}, n_{\mathrm{on}}, s_{\mathrm{traight}}, j_{\mathrm{unk}}$. With the

given 16 letters it is easy to construct a fold from a given name and reverse. One could go on by distinguishing also between the Chan loops, adding more vowels to the alphabet, etc.

[50] P. Flory, J. Chem. Phys. **10**, 51 (1941); H. Orland, C. Itzykson, and C. de Dominicis, J. Phys. (Paris) Lett. **46**, L353 (1985).

[51] To obtain the complete count of dense configurations for $Z=5$ according to the maximum number of neighbor criterion we must add for the number of elements $N=19$: 6164 from the $4\times1\times1$ box, for $N=29$: 4522202 from the $4\times2\times1$ box and for $N=29$: 7124170 and $N=31$: 8820464 from the $3\times3\times1$ box. Together with Table III this gives the exhaustive counts for N up to and including 35. For $Z=4$ we must add the numbers in Table IV.

[52] A. V. Finkelstein, A. M. Gunton, and A. Y. Badretdinov, FEBS Lett. **325**, 23 (1993).

[53] J. Bohr, H. Bohr, and S. Brunak, Europhys. News **27**, 50 (1996).

[54] H. Bohr and P. G. Wolynes, Phys. Rev. A **46**, 5242 (1992).

[55] B. Rost and C. Sanders, J. Mol. Biol. **232**, 584 (1993).

[56] Z. Huang, S. B. Prusiner, and F. E. Cohen, Folding and Design **1**, 13 (1996).

[57] P.-A. Lindgård and O. G. Mouritsen, Phys. Rev. Lett. **57**, 2458 (1986); Phys. Rev. B **41**, 688 (1990).

[58] In this reference we discuss briefly consequences of relaxing the equal length assumption for the secondary elements and the loops. Typically α helices and β strands are approximate cylinders, which are about 3 times as long and as wide. A packing in bundles of parallel cylinders on a tetragonal lattice with orthogonal lattice vectors $(a,b,c)\sim(a,a,3a)$ seems therefore a most likely choice for the parent state. The vertices in the projection on the xy plane then represent the secondary structures. The relevant numbers for this case are those for the vertices in the square xy plane. These were found by Chan and Dill [22] to be $N_{\mathrm{magic}}^{\mathrm{vertices}}(2D)=\{2,4,6,9,12,16,20,25,\ldots\}$. We argue that this is identical to the secondary magic numbers for the tetragonal phase: $s_{s_{\mathrm{tetragonal}}}=N_{\mathrm{magic}}^{\mathrm{vertices}}(2D)$. They deviate from our s_s only for $s_{s_{\mathrm{tetragonal}}}>16$. For large numbers of secondary structures it does not make sense to maintain the tetragonal packing (when the width of the pack gets much larger that the length of the secondary structures). Then it is important to consider the real 3D problem. We emphasize that the present interpretation of the 2D vertices and the identification of the corresponding magic number with s_s is new and not mentioned in Refs. [22, 39]. It is clear that for the tetragonal packing s_s is independent of whether the loops run along the x and y axis, respectively, or somehow diagonally—and whether the length is equal to, or larger than one unit. It is only the shape of the confinement of the vertices that matters. Consequently, the magic numbers, s_s, for bundlelike structures are very robust. These structures represent a subset of those we have considered in the model without the unequal length constraint and the magic numbers are identical. Finkelstein and Ptitsyn [5] previously considered the parallel packing problem. Again, they were concerned with the *native* state and discussed consequently the closest possible packing of α helices and β strands (even of unequal width). They did not discuss the more open *parent* state (for which is makes less sense to distinguish between the widths), and they did not discuss magic numbers.

[59] P. G. de Gennes, *Scaling Concept in Polymer Physics* (Cornell University Press, Ithaca, 1979).

[60] B. Li, N. Madras and A. D. Sokal, J. Stat. Phys. **80**, 661 (1995).

[61] C. Domb and M. E. Fisher, Camb. Philos. **54**, 48 (1957).

[62] The proposed hinge-force assisted folding is in fact a much more direct and reliable process than the corresponding defect assisted selection of variants, which occurs in ''trained'' shape memory alloys. Given the code for the hinge forces the names, i.e., fold classes can be directly predicted from the sequence information.

[63] E. Vives, T. Castán, and P.-A. Lindgård, Phys. Rev. B **53**, 8915 (1996).

[64] T. Castán and P.-A. Lindgård, Phys. Rev. B **40**, 5069 (1989).

[65] For the finite chain there is of course no phase transition in the strict sense, but it is replaced by a smoothed transition region.

[66] H. L. Richards, M. Kolesik, P.-A. Lindgård, P. A. Rikvold, and M. A. Novotny, Phys. Rev. B **55**, 11 521 (1997).

[67] E. I. Shakhnovich and A. V. Finkelstein, Biopolymers **28**, 1667 (1989).

[68] Unfortunately, Ptitsyn [2] often neglects the qualifying descriptions and calls the whole region of kinetic and equilibrium folding intermediates a molten globule, see, e.g., [6]. The word ''molten'' refers to a theoretical understanding of the phenomena [67] in terms of ''melted'' side chains; it does not refer to the fluctuations in the tertiary structure. In analogy to the Martensitic case, we would not call the parent state ''molten,'' although it is fluctuating.

[69] K. A. Dill and D. Shortle, Annu. Rev. Biochem. **28**, 5439 (1989).

[70] A. S. Landsberg and E. J. Friedman, Phys. Rev. E **54**, 3135 (1996).